Многомерная (матричная) формулировка. Метод наименьших квадратов.

Рассмотрим простейшую модель зависимости целевой переменной y от исходных данных x, Такая функция зависимости будет линейной:

$$f(w, x_i) = w_0 + w_1 x_{1i} + \ldots + w_k x_{ki}$$

Или, что тоже самое

$$y = w_0 + \sum_{i=1}^{m} w_i x_i$$

Если мы добавим фиктивную размерность $x_0 = 1$ для каждого наблюдения, тогда линейную форму можно переписать чуть более компактно, записав свободный член w_0 под сумму:

$$y = \sum_{i=0}^{m} w_i x_i = \vec{w}^T \vec{x}$$

Для матрицы признаков, у которой в строках находятся примеры из набора данных, нам необходимо добавить единичную колонку слева. Зададим модель следующим образом:

$$\vec{y} = X\vec{w} + \epsilon,$$

где

- $\vec{y} \in \mathbb{R}^n$ – целевая переменная;
- w – вектор параметров (или весов) модели;
- X – матрица признаков размерности n строк на $m+1$ столбцов (включая фиктивную единичную колонку слева) с полным рангом по столбцам: $\operatorname{rank}(X) = m + 1$;
- ϵ – случайная переменная, соответствующая случайной, непрогнозируемой ошибке модели.

Для каждого конкретного наблюдения, данное выражение имеет вид:

$$y_i = \sum_{j=0}^{m} w_j X_{ij} + \epsilon_i$$

Также на модель накладываются следующие ограничения (иначе это будет какая то другая регрессия, но точно не линейная):

- матожидание случайных ошибок равно нулю: $\forall i : \mathbb{E}[\epsilon_i] = 0$;

- дисперсия случайных ошибок одинакова и конечна, это свойство называется гомоскедастичностью: $\forall i : \operatorname{Var}(\epsilon_i) = \sigma^2 < \infty$;
- случайные ошибки не скоррелированы: $\forall i \neq j : \operatorname{Cov}(\epsilon_i, \epsilon_j) = 0$.

Оценка $\hat{w}_i$ весов w_i называется линейной, если

$$\hat{w}_i = \omega_{1i} y_1 + \omega_{2i} y_2 + \cdots + \omega_{ni} y_n,$$

где $\forall\, k\, \omega_{ki}$ зависит только от наблюдаемых данных X и почти наверняка нелинейно. Так как решением задачи поиска оптимальных весов будет именно линейная оценка, то и модель называется *линейной регрессией*. Введем еще одно определение. Оценка $\hat{w}_i$ называется несмещенной тогда, когда матожидание оценки равно реальному, но неизвестному значению оцениваемого параметра:

$$\mathbb{E}\left[\hat{w}_i\right] = w_i$$

Таким образом, наша задача состоит в том, чтобы подобрать такие коэффициенты w, при которых значения нашей аппроксимирующей функции $f(w, x_i)$ будут расположены максимально близко к значениям целевых показателей. Один из способов вычислить значения параметров модели является **метод наименьших квадратов** (МНК), который минимизирует среднеквадратичную ошибку между реальным значением зависимой переменной и прогнозом, выданным моделью. Функция оценки качества в таком случае примет следующий вид:

$$Err = \sum_{i=1}^{n} (y_i - f(x_i))^2 \rightarrow min$$

> До начала XIX в. не существовало общего математического метода решения системы уравнений, в которой число неизвестных меньше, чем число уравнений. Гауссу (1795) принадлежит первое применение метода, а Лежандр (1805) независимо открыл и опубликовал его под современным названием (Méthode des moindres quarrés). Лаплас связал метод с теорией вероятностей.
>
> Работы А. А. Маркова в начале XX века позволили включить метод наименьших квадратов в теорию оценивания математической статистики, в которой он является важной и естественной частью. См. Теорема Гаусса-Маркова

Таким образом, наша функция ошибки модели будет иметь вид:

$$\begin{aligned}
\mathcal{L}\left(X, \vec{y}, \vec{w}\right) &= \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \vec{w}^T \vec{x}_i\right)^2 \\
&= \frac{1}{2n} \left\| \vec{y} - X\vec{w} \right\|_2^2 \\
&= \frac{1}{2n} \left(\vec{y} - X\vec{w}\right)^T \left(\vec{y} - X\vec{w}\right)
\end{aligned}$$

Для решения задачи минимизации, воспользуемся следующими правилами матричного дифференцирования

$$
\begin{aligned}
\frac{\partial}{\partial x} x^T a &= a \\
\frac{\partial}{\partial x} x^T A x &= (A + A^T) x \\
\frac{\partial}{\partial A} x^T A y &= x y^T \\
\frac{\partial}{\partial x} A^{-1} &= -A^{-1} \frac{\partial A}{\partial x} A^{-1}
\end{aligned}
$$

Вычислим частную производную по параметрам:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \vec{w}} &= \frac{\partial}{\partial \vec{w}} \frac{1}{2n} \left(\vec{y}^T \vec{y} - 2 \vec{y}^T X \vec{w} + \vec{w}^T X^T X \vec{w} \right) \\
&= \frac{1}{2n} \left(-2 X^T \vec{y} + 2 X^T X \vec{w} \right)
\end{aligned}
$$

Приравнивая ее нулю, получим точное аналитическое решение для нашей задачи

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \vec{w}} = 0 \quad &\Leftrightarrow \quad \frac{1}{2n} \left(-2 X^T \vec{y} + 2 X^T X \vec{w} \right) = 0 \\
&\Leftrightarrow \quad -X^T \vec{y} + X^T X \vec{w} = 0 \\
&\Leftrightarrow \quad X^T X \vec{w} = X^T \vec{y} \\
&\Leftrightarrow \quad \vec{w} = \left(X^T X \right)^{-1} X^T \vec{y}
\end{aligned}
$$

Источники:

Открытый курс машинного обучения. Тема 4. Линейные модели классификации и регрессии / Хабр

Приводим уравнение линейной регрессии в матричный вид / Хабр

Линейная регрессия. Разбор математики и реализации на python / Хабр